산업융합연구 제18권 제4호 pp. 61-66, 2020
ISSN 2635-8875 / e-ISSN 2672-0124
https://doi.org/10.22678/JIC.2020.18.4.061

빅데이터를 활용한 타자의 출루 관련 경기력과 불쾌지수의 관계 분석 : 투구 수 유도와 출루율을 중심으로

김세민[1], 유강수[2*]
[1]전주교육대학교 컴퓨터교육과 외래교수, [2*]전주대학교 교양학부 부교수

Analysis of the Relationship between a Batter's Performance and Discomfort Index using Big Data: focusing on the Number of Pitches and On Base Percentage

Semin Kim[1], Kangsoo You[2*]
[1]Lecturer, Dept. of Computer Education, Jeonju National University of Education
[2*]Associate Professor, School of Liberal Arts, Jeonju University

요 약 최근 프로야구에서 데이터를 활용하여 경기, 시즌, 팀을 운영하려는 시도가 일반화 되고 있다. 이에 본 연구에서는 기상 응용 데이터인 불쾌지수와 같은 경기 외적인 요소를 야구 경기 기록을 수집하고 출루율과 투구 수 유도와의 관계를 분석하였으며 이를 3차 기록으로 정의하여 연구를 수행하였다. 불쾌지수가 75이상일 때 투수의 투구 수 유도가 많이 되었으며, 불쾌지수가 69.9 이하일 때는 출루율이 높게 나왔으나, 불쾌지수가 70 이상 75미만일 때는 타자의 출루 관련 경기력이 가장 저조한 것으로 나타났다. 연구 결과를 통하여 불쾌지수와 타자의 출루율과 투구 유도 수는 관계가 있으며, 투수의 경기력과 관계있을 가능성이 높다고 유추할 수 있었다. 본 연구를 통하여 1차 기록이라 정의하는 누적·비율기록과 2차 기록이라 정의하는 세이버메트릭스에 이어서 경기 외적인 데이터를 연계하는 3차 기록으로 정의할 수 있는 가능성을 볼 수 있었다.

키워드 : 빅데이터 분석, 스포츠 데이터, 한국프로야구, 불쾌지수, 경기력

Abstract Recently, attempts have been made to use data to operate games, seasons, and teams in professional baseball. Therefore, in this study, we collected baseball game records and analyzed the relationship between on-base rate and pitching count induction, and this was defined as the third record for non-game factors such as discomfort index, which includes the weather application data. When the discomfort index was over 75, the pitcher's induction of pitching was high, and when the discomfort index was less than 69.9, the on-base rate was high, but when the discomfort index was 70 or more and less than 75, the batter's on-base performance was the lowest. Through the results of the study, it could be inferred that the discomfort index, the batter's on-base rate, and the number of induction pitches are related, and that it is highly likely to be related to the pitcher's performance. Through this study, we could see the possibility of defining a cumulative/ratio record defined as the primary record and a saver metric defined as the secondary record, and a third, tertiary record linking data outside the game.

Key Words : Big Data Analysis, Sports Data, KBO League, Discomfort Index, Performance

1. 서론

최근 빅데이터 기술이 주목받음에 따라 스포츠 분야에서도 데이터를 활용하려는 시도가 일반화 되고 있다. 야구의 세이버메트릭스(Sabermetrics)와 농구의 APBR메트릭스(APBRmetrics)는 기존의 클래식 기록

*Corresponding Author : Kangsoo You(gsyou@jj.ac.kr)
Received July 20, 2020 Revised August 14, 2020
Accepted August 21, 2020 Published August 31, 2020

과 함께 미디어에 게시되거나 일반 팬들도 자주 열람하는 데이터가 되었다. 또한 트랙맨(Trackman) 시스템과 같은 영상처리 기술의 발달로 인하여 영상을 분석하여 투구 시 투수의 팔 각도, 선수가 운동장을 달리는 속도, 타구의 발사각도와 궤적, 투구 시 공의 회전 수 등을 분석하기까지 이르렀다. 이에 따라 스포츠 분야에서 다양한 형태의 데이터가 많이 발생하게 되었고 다양한 수요가 창출되고 있다[1-3].

특히 야구는 데이터가 많이 창출되는 종목이다. 경기 시간이 3시간 내외로 긴 편이며, 프로 스포츠 중 가장 경기 수가 많으며, 가장 많은 선수가 출전한다. 또한 1차적인 단순한 누적 기록과 비율 기록을 창출하는 방법도 공격, 수비, 주루, 투구 등 명확하게 구분된 다양한 분야가 존재하고, 서로 밀접하게 관련된 기록들이 창출된다. 최근에는 효율성을 측정하기 위하여 과학적인 통계방법을 활용한 2차 기록인 세이버메트릭스를 활용하여 경기 운영, 시즌 운영, 팀 운영을 하기에 이르렀다[4-6].

대한민국은 다양하고 뚜렷한 기후를 가지고 있다. 또한 리그 진행 기간이 상당히 길다. 또한 대한민국의 야구장은 고척 스카이돔을 제외하고는 천장이 없는 야외 구장이다. 따라서 기후는 야구 경기에 영향을 주는 중요한 요소이다. 야구는 투구와 타격으로 경기가 시작된다. 투수가 타자를 아웃시키기 위하여 필요한 투구 수는 1개 이상이며, 심한 경우에는 10개 이상의 투구가 필요할 수 있다. KBO리그에서는 '용규놀이'라는 유행어가 있다. 현 한화 이글스의 이용규 선수는 끊임없이 파울을 만들어서 상대 투수를 괴롭히기로 유명하다. 투수는 계속된 파울로 많은 횟수의 공을 던져야 하며, 타자에게 출루를 허용하면 심리적으로나 체력적으로 손해를 보게 되며, 설사 타자를 아웃시키더라도 체력이 더욱 고갈되어 강판당하는 시간이 앞당겨진다. 따라서 '용규놀이'를 잘하는 타자는 팀 기여도가 높다. 설상가상으로 더운 여름에는 상대 투수에게 심리적 불쾌감을 줄 수 있다. 이에 본 연구에서는 기상 응용 데이터인 불쾌지수를 활용하여 투구 수와 출루율과의 관계를 분석하고자 한다[7, 8].

본 연구를 위하여 필요한 조작적 정의는 1차 기록인 누적기록과 비율기록, 2차 기록인 세이버메트릭스에 이어, 3차 기록은 경기 외부 조건을 활용한 기록이라 정의한다.

2. 선행연구

2.1 불쾌지수

불쾌지수는 대한민국의 기상청에서 제공하는 기상자료개방포털(http://data.kma.go.kr) 공식 홈페이지를 통하여 열람할 수 있다. 우리나라에서는 여름철에 해당하는 6월에서 9월 사이에 불쾌지수를 측정하여 발표하고 있다. 불쾌지수는 우리 생활 속에서 가장 많이 언급하는 기상 응용 데이터이다. 야구는 야외 스포츠이자 멘탈 스포츠이므로 불쾌지수는 중요한 변수가 될 수 있음에 주목하였다. 불쾌지수는 70 이상에서 10% 이상인 사람이 불쾌감을 느끼고, 75 이상에서 50% 이상인 사람이 불쾌감을 느끼며, 80 이상에서 대부분의 사람이 불쾌감을 느낀다고 하지만 사람마다 기준이 다르므로 절대적이지 않으며 상황마다 다르다[9, 10].

2.2 스포츠와 빅데이터

야구는 수많은 데이터를 양산하는 종목이다. 따라서 타격, 수비, 주루, 투구 등의 행위에서 파생되는 각종 기록들이 현장에서 바로 기록되며, 데이터 수집, 저장, 분석, 처리, 배포 등이 예로부터 활발하게 진행되어왔다[11]. 과거에는 기록지에 기록하여 보관하는 형태로 진행되어왔으나 최근에는 정보통신기술의 발전으로 야구 관련 공식 홈페이지와 각종 포털 사이트에 저장되고 있다[2].

아울러 영상처리 및 방송 기술을 활용하여 영상정보를 숫자로 변환하여 활용하기도 하며, 최근에는 AI를 이용하여 스트라이크와 볼을 구분하는 의견까지 개진되고 있다. 여러 가지 기술을 경기 운영과 시즌 운영뿐만 아니라 구단 운영에까지 활용하게 되면서 프로야구 구단들은 데이터 과학자들을 초빙하여 활용하고 있다[12-13].

2.3 관련 야구 기록

본 연구에서 활용하는 기록인 출루율은 타자가 타석에 들어서서 출루를 얼마나 하는지 측정한 기록이다. 주요 프로야구리그 중에서는 최초로 KBO가 프로야구출범 원년부터 출루율을 공식 기록으로 인정하고 출루율 1위 선수를 시상하였다. 미국 메이저리그(MLB)에서는 1984년에 공식 기록으로 채택하였고 일본 프로야구(NPB)에서도 공식 기록으로 인정하였다.

출루율은 개인기록이기도 하지만 팀에 더욱 도움

이 되는 기록이다. 야구는 주자가 홈을 다시 밟아야 득점이 되므로 반드시 출루가 이루어져야 한다. 따라서 최근에는 타율보다 출루율을 중요시하는 야구 전문가도 많아지는 추세이다. 출루율을 계산하는 공식은 아래와 같다[14].

출루율 = (안타 + 볼넷 + 몸에 맞은 공) / (타수 + 볼넷 + 몸에 맞은 공 + 희생 플라이)

투구 수는 투수의 기록이기도 하고 타자의 기록이기도 하다. 선발투수도 100개 이상의 공을 투구하는 것이 쉬운 일이 아니다. 따라서 최근에는 공을 많이 던지게 하는 타자도 각광을 받는 시대가 되었다.

3. 연구 방법

3.1 연구 대상

본 연구에서는 불쾌지수가 타자의 출루율과 투구 수 유도에 미치는 영향을 분석하기 위하여 야구장이 있는 지역의 불쾌지수를 수집하였다. 불쾌지수는 기상자료개방포털(https://data.kma.go.kr/cmmn/main.do)에서 수집하였다.

또한 2019년에 진행된 KBO 리그에서 규정 타석에 진입한 선수를 대상으로 출루율과 투구 유도 수를 수집하였다. 이를 위하여 KBO 공식 홈페이지(http://www.koreabaseball.com)와 스탯티즈 공식 홈페이지(http://www.statiz.co.kr)를 참고하여 데이터를 수집하였다. 규정 타석에 진입한 선수만을 대상으로 한 이유는 2019년 내내 꾸준한 활약을 하였을 가능성이 높기 때문이다. 규정 타석에 진입하지 않은 선수들은 꾸준히 출전한 데이터가 쌓여있지 않을 가능성이 높다. 예를 들어 불쾌지수가 측정되는 6월에서 9월까지만 활동한 선수도 있을 수도 있고, 그 반대의 경우가 있을 수 있다. 이들은 1년 내내 꾸준히 출전한 선수와 같은 조건이라고 보기 힘들기 때문이다.

불쾌지수는 불쾌감을 거의 느끼지 않는 69.9이하, 10% 이상의 사람이 불쾌감을 느끼는 70에서 74.9, 50% 이상의 사람이 불쾌감을 느끼는 75이상 등의 3단계로 나누어 데이터를 분류하였다.

3.2 데이터 분석

본 연구를 위하여 활용한 빅데이터 분석 방법은 SAS(Statistical Analysis System)사에서 제안한 SEMMA(Sampling Exploration Modification Modeling Assessment) 모형이다. SEMMA모형에 의하여 불쾌지수와 규정타석에 진입한 타자의 데이터를 수집하고, 이를 데이터베이스에 저장하거나 CSV 파일로 수집한 후, 쓸모 없는 데이터를 걸러내거나 분석을 위하여 데이터를 재배치하는 등 데이터 전처리 과정을 거쳤다.

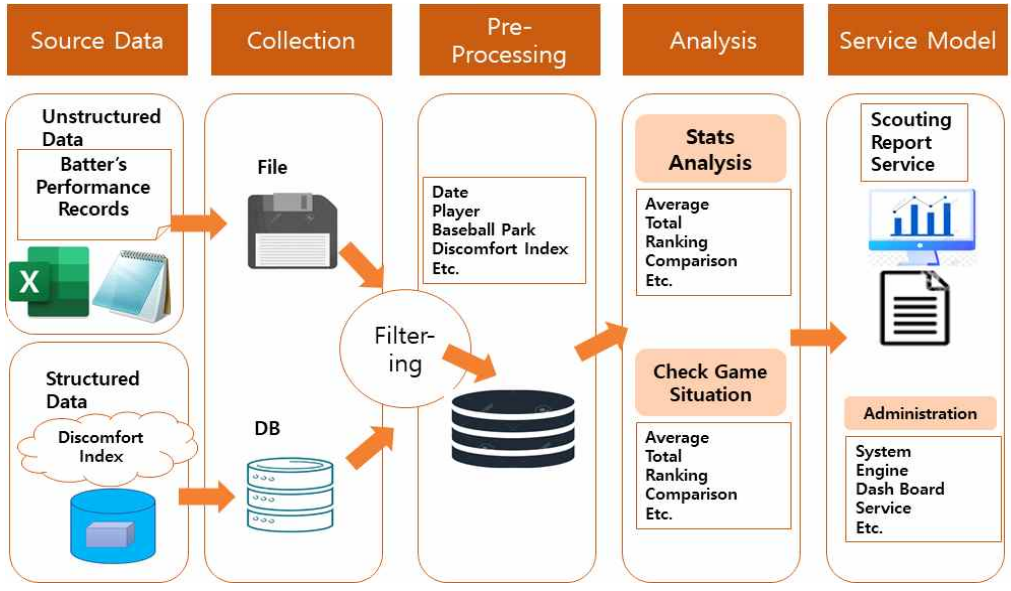

Fig. 1. Data Analysis Process

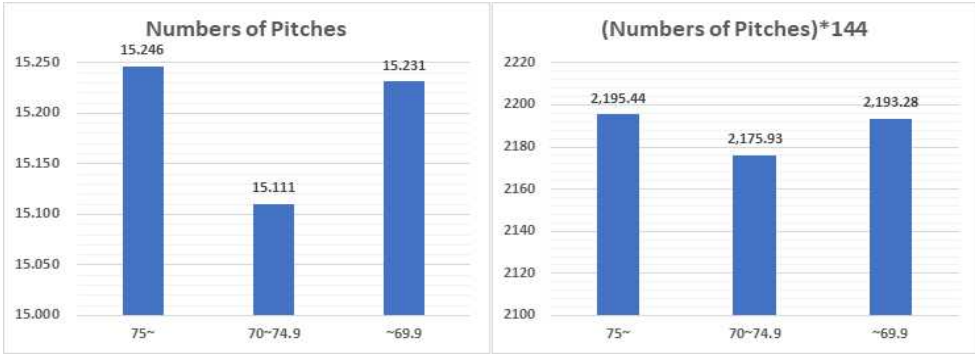

Fig. 2. Numbers of Pitches

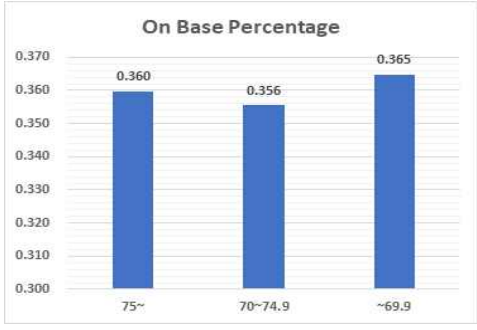

Fig. 3. On Base Percentage

분석과정에서는 규정 타석에 진입한 타자들의 출루율과 투구 유도 수 데이터를 평균, 합계, 비교분석, 순위 등으로 분석하고 그래프를 통하여 시각화 하였다. 본 연구를 통하여 향후 연구과제나 비즈니스로의 연결을 위하여 보고서 작업과 시각화 작업으로 스카우팅 레포트(Scouting Report)를 작성할 수 있다. 이를 위한 구체적인 과정은 Fig. 1과 같다.

3.3 연구의 제한점

본 연구에서 진행한 과정과 결과를 원활하게 도출하기 위하여 다음과 같은 몇 가지 제한점을 두었다.

첫째, 본 연구에서는 불쾌지수를 대상으로 하였다. 따라서 2019년 6월에서 9월까지 진행된 야구 경기 데이터를 대상으로 분석하였다.

둘째, 본 연구는 KBO리그의 2019년 기록만을 대상으로 하였기 때문에 해당 선수의 경기력에 대하여 완전한 일반화를 할 수는 없다. 그러나 일부 선수들은 2019년을 기점으로 경기력이 한단계 상승하거나 에이징 커브(Aging Curve)로 인하여 하락하는 계기가 되는 시기가 될 수 있으므로 추적 연구를 한다면 의미 있는 데이터로 발전시킬 수는 있다.

셋째, 2019년 기록만을 대상으로 하였기 때문에 2019년이 KBO 리그를 대표하는 연구 결과가 될 수 없다. 왜냐하면 타고투저나 투고타저라는 용어가 있듯이 같은 출루율 0.400이더라도 리그 전체 10위권 기록일 수도 있고, 리그 전체 20위권 기록일 수도 있기 때문이다.

4. 연구 결과

4.1 투구 유도 수

KBO 리그는 팀당 144경기를 진행한다. 따라서 경기 당 투구 유도 수와 144경기 당 투구 유도 수를 Fig. 2와 같이 계산하여 분석하였다. 2019년 KBO 리그의 규정 타석에 진입한 타자들의 경기 당 투구 유도 수는 불쾌지수 69.9 이하에서는 15.231구를 기록하고, 불쾌지수 70~74.9에서는 15.111구를 기록하였며, 불쾌지수 75 이상에서는 15.246구를 기록하였다. 144경기로 환산한 투구 유도 수는 불쾌지수 69.9 이하에서는 약 2,193구를 기록하였고, 불쾌지수 70~74.9에서는 2,175구를 기록하였으며, 불쾌지수 75 이상에서는 2,195구를 기록하였다.

4.2 출루율

본 연구에서 2019년 KBO 리그의 규정 타석에 진입한 타자들의 출루율을 분석한 결과로는 Fig. 3과 같

으며, 불쾌지수 69.9 이하에서는 출루율 0.365를 기록하였고, 불쾌지수 70~74.9에서는 출루율 0.356을 기록하였으며, 불쾌지수 75 이상에서는 출루율 0.360을 기록하였다.

5. 논의

연구 결과에서 분석된 결과를 토대로 종합적으로 다음과 같이 논할 수 있다.

첫째, 타자들의 투구 유도 수 기록은 불쾌지수 75 이상일 때 가장 높게 나타났다. 불쾌지수는 여름에만 측정되는 데이터이며 온도와 습도가 높을수록 높게 나타난다. 투수는 마운드에 계속 공을 던지면서 서있게 되므로 체력 소모가 크다. 따라서 덕 아웃이라는 상대적으로 쾌적한 환경에서 대기하고 있다가 타석에 등장하는 타자보다 계속 마운드에서 계속 투구를 진행하는 투수에게 체력적으로 더욱 불리하게 작용할 가능성이 높다. 투수의 체력이 떨어지면 공을 쥐는 악력이 감소하므로 제구력에 영향을 미치고 공을 던지는 개수가 많아진다는 것은 야구계에서 널리 통용되는 통설이다[14]. 따라서 불쾌지수가 높을수록 타자가 많은 공을 유도할 가능성이 높다고 할 수 있다.

둘째, 타자들의 출루율은 불쾌지수 69.9 이하에서 가장 높게 나타났다. 출루율 기록은 안타가 포함된다. 타자가 친 타구가 안타가 되려면 내야보다는 외야로 나갈수록 확률이 높은 편이다. 또한 불쾌지수가 낮으려면 온도와 습도가 낮아야 한다. 습도가 낮을수록 타구의 비거리는 길어진다[15]. 따라서 불쾌지수가 낮은 환경이 낮은 습도로 인하여 타자에게 더 유리하게 작용된다고 볼 수 있다.

셋째, 불쾌지수 70~74.9의 환경에서는 타자의 기록이 가장 저조하였다. 이는 높은 불쾌지수로 인하여 투수의 체력에 영향을 받거나, 낮은 불쾌지수로 인하여 타구의 비거리의 영향을 받지 않고, 투수가 경기력을 발휘한 결과로 논의할 수 있다.

6. 결론 및 제언

본 연구에서는 빅데이터 분석 방법을 활용하여 불쾌지수라는 기상 응용 데이터와 타자의 출루율과 투구 유도 수를 수집하였다. 이를 근거로 불쾌지수가 타자들의 출루 관련 경기력에 미치는 영향을 분석하였다. 출루 관련 경기력은 투구 수 유도와 출루율을 분석하였다. 우리나라는 동위도의 지역에 비하여 유독 여름이 무덥고 습한 편이다. 이에 불쾌지수가 높은 환경에서는 높은 습도와 높은 온도로 인하여 투수들의 체력 저하가 되므로 타자들이 공을 오래 지켜본다면 투구 수 유도를 이루어낼 수 있으며, 불쾌지수가 낮은 환경에서는 낮은 습도로 인하여 타자들의 비거리가 증가하여 안타 생산이 증가하므로 출루율 증가에 보탬이 된다는 것을 알 수 있었다. 따라서 야구선수들의 경기력에 많은 영향을 끼치며, 투수의 기록도 타자의 출루 관련 경기력과 관계가 있다는 것도 논할 수 있었다.

또한 본 연구의 결과를 통하여 기존의 1차 기록인 누적기록과 비율기록, 2차 기록인 세이버메트릭스에 이어서, 3차 기록으로 외부 환경 요인과 경기 데이터를 적용하여 활용할 수 있는 가능성을 확인하였다.

향후 연구과제로는 투구, 주루, 수비 기록과 타격 기록을 연계하여 분석하는 것이다.

REFERENCES

[1] S. M. Kim. (2020). *The effect of daily average temperature on the batter's performance in baseball game : focused on big data analysis.* Master's Thesis. The Graduate School of Hoseo University, Asan, Chungnam.

[2] J. Y. Hong. (2019). *The effect of golf pre shot routine on club and ball data.* Master's Theis. The Graduate School of Choongang University, Seoul.

[3] S. H. Lee & H. J. Choi. (2019). The analysis of pitching result according to the velocity and pitch of pitcher in that case of full-counting on Major League Baseball(MLB). *The Korea Journal of Sports Science, 28(3),* 973-981. DOI : 10.35159/kjss.2019.06.28.3.973

[4] S. M. Kim & K. S. You. (2020). The effect of daily average humidity on pitcher's stats of a strike-out : focused on high rankers of winning, hold and save. *Journal of Industrial Convergence, 18(1),* 65-71. https://doi.org/10.22678/JIC.2020.18.1.065

[5] Y. H. Kim. (2020). *Analysis of football tactics and formation patterns based on big data analysis*. Master's Thesis. The Graduate School of Choongang University, Seoul.

[6] Rein, R., & Memmert, D. (2016). Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *SpringerPlus, 5(1)*, 1-13.

[7] J. W. Lee & C. H. Lee. (2019). A study on the analysis of news data for the improvement of local flower festival. *Journal of Industrial Convergence*, *17(4)*, 33-38.
DOI : 10.22678/JIC.2019.17.4.033

[8] J. H. Gang. (2020). *Who is the number one player in "Yongkyu Play" this season?*. Hankookilbo(Website). https://www.hankookilbo.com/News/Read/A2020070914160005513?did=NA

[9] I. J. Jeon & K. Y. Chung. (2009). Life weather index monitoring system using wearable based smart cap. *The Journal of the Korea Contents Association, 9(12)*, 477-484.

[10] J. M. Kim, M. S. Kim & K. N. Kim. (2014). Crime prediction model based on meteorological changes and discomfort index. *Convergence Security Journal, 14(6)*, 89-95.

[11] J. T. Lee. (2015). Long term trends in the Korean professional baseball. *Journal of the Korean Data & Information Science Society, 26(1)*, 1-10.

[12] H. S. Seok & Y. J. Lee. (2019). Ontology-based IoT context information modeling and semantic-based IoT mashup services implementation. *Journal of the KIECS, 14(4)*, 671-678.
DOI : 10.13067/JKIECS.2019.14.4.671

[13] Herm, S., Callsen-Bracker, H. M. & Kreis, H. (2014). When the crowd evaluates soccer players' market values: Accuracy and evaluation attributes of an online community. *Sport Management Review, 17(4)*, 484-492.

[14] J. W. Song. (2020). Ryu Hyun-jin and Kim Gwang-hyun left big homework as expected, Sisa Journal(Online), http://www.sisajournal.com/news/articleView.html?idxno=203191.

[15] Y. R. Ko. (2016). [Nexen Column], *Temperature and humidity, hidden effect of Gocheok Skydome*. 10 Baseball club's column(Online), https://sports.news.naver.com/news.nhn?oid=525&aid=0000000070.

김세민(Semin Kim) [종신회원]

- 2006년 2월 : 우석대학교 컴퓨터교육과(교육학석사)
- 2009년 8월 : 공주대학교 컴퓨터교육학과(교육학박사수료)
- 2018년 8월 : 한밭대학교 정보통신공학과(공학박사)
- 2020년 2월 : 호서대학교 스포츠과학대학원 야구학과(체육학석사)
- 2008년 3월 ~ 현재 : 전주교육대학교 외래교수
- 관심분야 : 스포츠데이터과학, 빅데이터, 소프트웨어교육, 메이커교육
- E-Mail : imsil303@hotmail.co.kr

유강수(Kangsoo You) [종신회원]

- 2005년 8월 : 전북대학교 영상공학과(공학박사)
- 1996년 3월 ~ 2006년 8월 : 전주대학교 교양학부 객원교수
- 2006년 9월 ~ 현재 : 전주대학교 교양학부 교수
- 관심분야 : 스포츠데이터과학, 영상처리, 컴퓨터비전, 소프트웨어교육
- E-Mail : gsyou@jj.ac.kr